TUTORIAL BY H4

COME COPIARE INTERAMENTE UN SITO WEB ELUDENDO LE RESTRIZIONI IMPOSTE DAL ROBOTS.TXT GRAZIE A HTTRACK

Uno dei problemi principali di quando si cerca di copiare un sito web, è che non è possibile a causa di un divieto imposto dal file Robots.txt.

Per informazioni su cos'è un file Robots.txt rimando alla pagina:

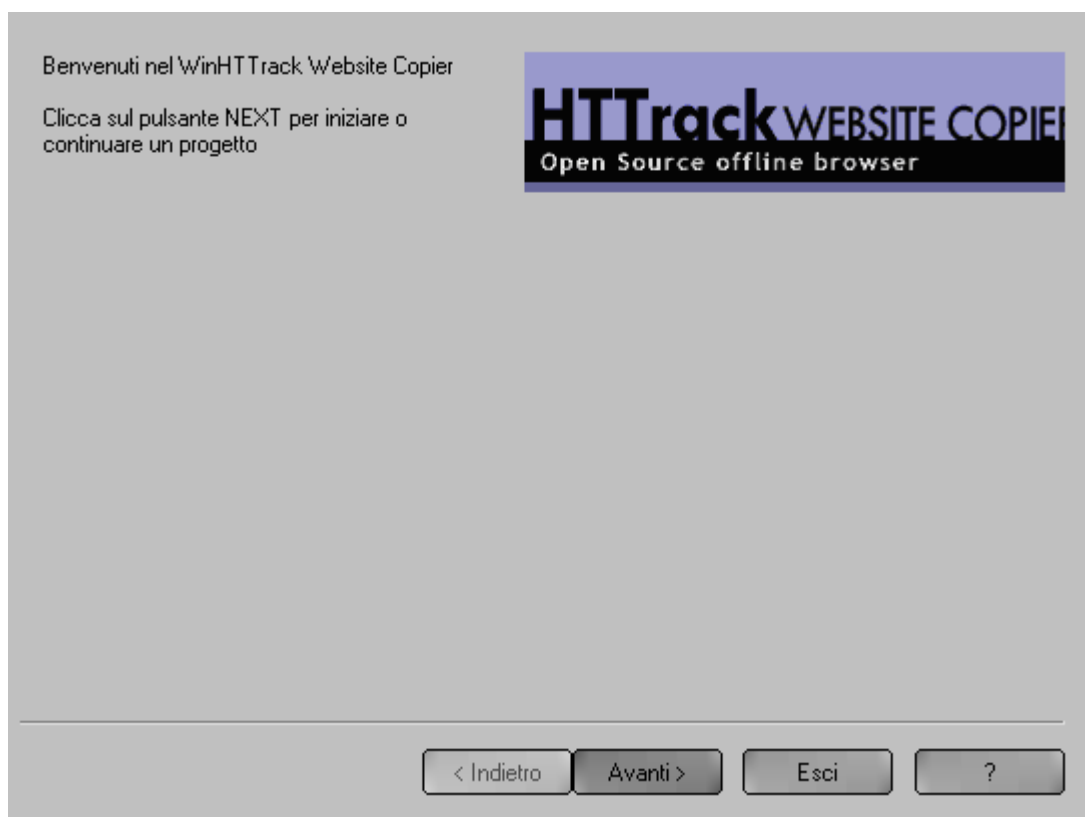
http://it.wikipedia.org/wiki/Robots.txt

Iniziamo ad analizzare i metodi per scaricare un Sito Web grazie al tool gratuito HTTrack disponibile in multipiattaforma sia per Windows che per Linux.
La differenza sostanziale sulle due piattaforme è che su Windows è stata sviluppata una GUI per rendere il lavoro più semplice, al contrario del lavoro di shell richiesto per Linux.

Lo potete scaricare questo indirizzo:

http://www.httrack.com/

Procedura passo passo di utilizzo"WinHTTrack.exe"(Interfaccia GUI per HTTrack per Win)

Aprire "**WinHttrack.exe**"

Avanti >

Scegliere il nome del progetto, la categoria e la destinazione di salvataggio del progetto.
Avanti >

Scegliere la tipologia di azione da parte del programma (Scarica il sito permette di scaricare interamente il Sito Web)

Adesso un passo importante, dobbiamo scegliere le opzioni per il download del sito, queste sono da scegliere con attenzione, necessarie per la buona riuscita del lavoro. *Per accedere al pannello di controllo delle opzioni basterà cliccare su **Definisci opzioni** della finestra precedente.*

La mia attenzione ricade sull'opzione Identità del Browser perchè scegliere tra le varie opzioni di questa sezione è importante per permettere il download delle pagine.
Per capire il motivo di questa enfasi su quella opzione bisogna sapere come funziona un web downloader.
HTTrack come molti web downloader accede ad una pagina, parsifica tutto il codice HTML alla ricerca di link ad altre pagine (grazie ad uno spider) e salva una copia di ogni pagina che visita nella cartella desiderata; fin qui nessun problema, ma dobbiamo aggiungere il fatto che il Robots.txt impone delle restrizioni per l'accesso alle pagine web del sito da parte degli spider.
Essendo il Robots.txt uno strumento creato dal webmaster per marcare quei file e quelle directory di un sito web che non si vuole rendere accessibili agli spider dei motori di ricerca, possiamo facilmente verificare che nel nostro caso, nel file Robots.txt localizzato nell'indirizzo:

http://*.*******.**/Robots.txt**

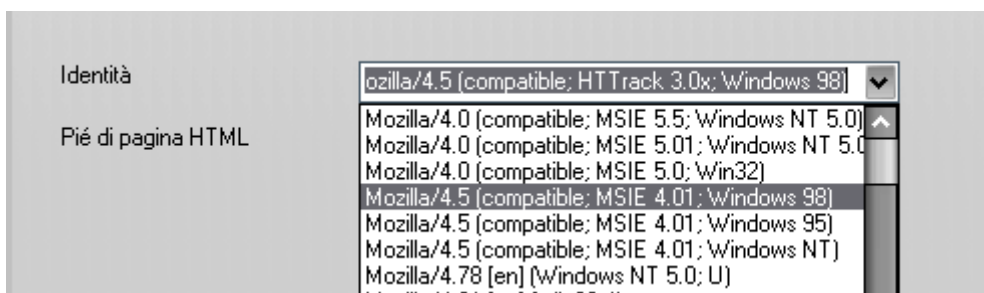
troviamo una stringa del genere:

```
#Aggiunto il 03 agosto 2002
User-agent: WinHTTrack
Disallow: /
```

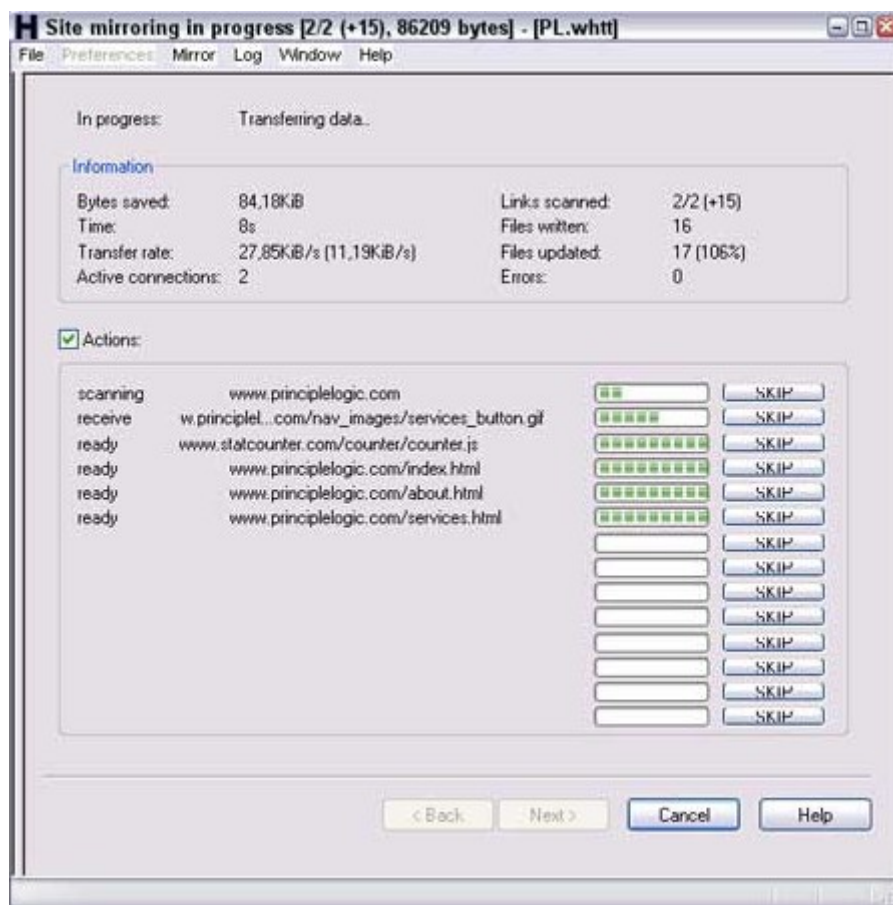
Possiamo ben capire che lo User-Agent WinHTTrack non ha il permesso di accedere a nessuna pagina del sito.

Come possiamo ovviare a questo problema? Semplicemente cambiando identità al nostro User-agent, dandogliene una che non è presente nel file Robots.txt.

Con il tool WinHTTrack nella sezione Identità del Browser possiamo cambiare identità al nostro User-Agent facendo in modo che appaia non più come:

User-agent: WinHTTrack ma come User-agent: WinHTTrack

Secondo la mia esperienza quello selezionato funziona su una marea di siti internet, perchè non viene mai bloccato, quindi scegliete quello.

Le altre opzioni non le tratto perchè sono superflue a causa della variabilità che hanno a seconda del tipo di banda a disposizione, spazio... ecc.

Avanti e fine. Vedremo poi iniziare il download e la scansione da parte degli spiders delle diverse pagine.

Diversa la situazione con le impostazioni tramite Shell. Di seguito sono riportate passo passo le istruzioni da seguire per lo stesso scopo illustrato nel punto precedente.

Aprite "**httrack.exe**" o diversa estensione a seconda del SO.

Inserite il nome del progetto:

```
Welcome to HTTrack Website Copier (Offline Browser) 3.42+htsswf+htsjava
Copyright (C) Xavier Roche and other contributors
Note: You are running the commandline version,
run 'WinHTTrack.exe' to get the GUI version.
To see the option list, enter a blank line or try httrack --help

Enter project name :Prova2
```

Inserire la cartella di destinazione:

```
Base path (return=./websites/) :c:/prova2
```

Inserire l'Url:

```
Enter URLs (separated by commas or blank spaces) :http://css.html.it
```

Scegliere l'azione da compiere:

```
Action:
(enter) 1       Mirror Web Site(s)
        2       Mirror Web Site(s) with Wizard
        3       Just Get Files Indicated
        4       Mirror ALL links in URLs (Multiple Mirror)
        5       Test Links In URLs (Bookmark Test)
        0       Quit
: 1
```

Premiamo due volte invio per non settare le impostazioni Proxy e le Wildcards:
(return=none) significa che se premete invio la scelta è "nessuno"

```
Proxy (return=none) :

You can define wildcards, like: -*.gif +www.*.com/*.zip -*img_*.zip
Wildcards (return=none) :
```

Definiamo le opzioni:

```
You can define additional options, such as recurse level (-r<number>), separed b
y blank spaces
To see the option list, type help
Additional options (return=none) :
```

Per visualizzare tutte le opzioni digitare help e poi premere invio.
Nel nostro caso dobbiamo settare l'identità del browser, seguiamo la seguente linea di comando

-F "Mozzilla/4.5"

```
You can define additional options, such as recurse level (-r<number>), separed b
y blank spaces
To see the option list, type help
Additional options (return=none) :F Mozzilla/4.5
```

Avremo come risposta la wizard commandline precompilata:

```
---> Wizard command line: httrack http://css.html.it  -O "c:\prova3/oipgjreipojg
r"  -%v -F "Mozilla/4.5"
```

Ci verrà richiesta la conferma del mirroring:
Digitiamo Y

```
Ready to launch the mirror? (Y/n) :y
```

Attendiamo la fine del processo:

```
Mirror launched on Sat, 19 Apr 2008 01:03:33 by HTTrack Website Copier/3.42+htss
wf+htsjava [XR&CO'2007]
mirroring http://css.html.it with the wizard help..
```

Verrà concluso quando la shell si chiuderà da sola.

Enjoy.